UNIVERSITA' DEGLI STUDI DI PADOVA

DIPARTIMENTO DI SCIENZE ECONOMICHE ED AZIENDALI "M. FANNO"

DIPARTIMENTO DI SCIENZE STATISTICHE

CORSO DI LAUREA IN ECONOMIA

PROVA FINALE

SENTIMENT ANALYSIS PER LE IMPRESE: STRUMENTI E APPLICAZIONI

RELATORE:

CH.MO PROF. TOMMASO DI FONZO

LAUREANDO: RICCARDO ENNIO

MATRICOLA N. 1088823

ANNO ACCADEMICO 2016 – 2017

INDICE

INTRODUZIONE

La comunicazione con i clienti e in particolare la gestione delle informazioni, soprattutto se provenienti dalla Rete, rappresenta un punto cardine nello svolgersi dell'attività manageriale di un'azienda. Per le imprese è infatti fondamentale ottenere in tempo reale informazioni sull'apprezzamento dei propri prodotti e servizi, sul tipo di emozioni che i clienti associano al brand e sul tenore di consigli e opinioni che i consumatori si scambiano fra loro. Detenere queste conoscenze diventa un vantaggio competitivo nei confronti dei *competitor* sul mercato poiché permette di adattare la propria offerta alle necessità e alle preferenze espresse direttamente dai clienti personalizzandone le caratteristiche.

Non a caso, molte imprese hanno deciso di sviluppare internamente un Sistema Informativo di Marketing (SIM), che consiste in un apparato costituito da risorse umane, tecnologie e procedure destinato alla valutazione e allo sviluppo del fabbisogno di indicazioni necessarie ad orientare i responsabili delle decisioni di marketing a scegliere al meglio. Aspetto cruciale per tale sistema è il poter disporre di informazioni affidabili che possono essere usate in modo efficace per generare e convalidare dati su clienti e mercati (Kotler, Armstrong, Ancarani e Costabile, 2015).

Negli ultimi anni l'utilizzo diffuso del *world wide web*, come sede di scambio di informazioni e di opinioni, è aumentato a dismisura. Questo tipo di andamento sembra non accennare a diminuire, anzi il fenomeno web e la conseguente diffusione di strumenti per accedervi sembrerebbe incrementare la sua portata anno dopo anno; alcuni studi affermano che nel 2020 ci saranno più di 75 miliardi di dispositivi collegati a Internet, mentre si prevede che la popolazione mondiale raggiunga gli 8 miliardi, giungendo ad un numero di dispositivi per persona di poco inferiore a 10.

È dunque immediato comprendere come i principali social media, quali ad esempio *Twitter*, *Facebook* e *Google +*, detengano un potere informativo ineguagliabile. Ogni giorno milioni di persone accedono a queste piattaforme producendo un'enorme quantità di commenti ed esprimendo le proprie opinioni sui più disparati argomenti, dagli ultimi avvenimenti sportivi alla politica, dai programmi televisivi alle preferenze in termini di prodotto. Questo bacino informativo è l'oggetto di studio cui si riferiscono le pratiche di *sentiment analysis*: traducendo i commenti e, più in generale, i testi creati attraverso l'uso dei social network, in dati numerici statisticamente e quantitativamente rilevanti, è possibile, ad esempio, verificare quale tipo di

predisposizione emotiva è più diffusa nei confronti di un prodotto o di una campagna pubblicitaria (per citare solo alcune tra le applicazioni di interesse per questo lavoro).

Il presente elaborato si pone come obiettivo quello di analizzare come i nuovi sistemi di analisi testuale, in costante e rapida evoluzione, applicati all'ambito dei social network, possano influenzare le strategie e l'offerta delle aziende, dalle scelte di marketing a quelle di produzione, o, ancora, al servizio clienti, arrivando in certi casi a prevedere con un discreto anticipo le tendenze e la predisposizione emotiva dei consumatori nei confronti dell'azienda. Questo intento verrà perseguito attraverso una rassegna delle principali tecniche di analisi del *sentiment* presenti ad oggi in letteratura.

L'ambito dell'analisi testuale riferita a volumi consistenti di dati eterogenei, sia strutturati che non strutturati, i *Big Data*, attraverso software e programmi informatici, è vasto e complesso proprio a causa della sua natura dinamica e in costante evoluzione. Il tema porta alla luce non pochi spunti di riflessione che coinvolgono campi come l'informatica, la statistica, l'economia e la politica. In questo lavoro si è preferito concentrarsi maggiormente sulle implicazioni socio-economiche che questo argomento pone, evitando la descrizione in profondità del funzionamento degli algoritmi e dei software che li implementano sotto i punti di vista dei modelli statistici e degli strumenti informatici utilizzati; questa scelta è dovuta ad una maggiore consapevolezza delle tematiche economiche connesse al tema della *sentiment analysis* rispetto alle questioni di tipo statistico e informatico dello stesso, non trattate nel corso di studi frequentato dallo scrivente.

Sulla base di questi assunti, l'elaborato è organizzato nel modo seguente.

- Nel Capitolo 1 viene delineata una definizione della locuzione *sentiment analysis* alla luce della letteratura di settore, analizzando anche il concetto largamente utilizzato di *opinion mining*. Segue una breve spiegazione del contesto in cui operano gli strumenti di analisi testuale riferiti a tali modelli, in particolare i social network; si analizzerà come gli ambienti online siano diventati un punto di riferimento per il reperimento di informazioni rilevanti mediante una breve panoramica sui principali motivi per i quali il mondo dei *social* rappresenti ad oggi il campo da gioco più importante per le analisi del *sentiment*. Si forniranno dati aggiornati riguardo l'utilizzo della rete nel mondo. In chiusura di Capitolo si evidenzieranno quelli che, secondo gli studi più recenti, possono essere considerati

i principali motivi di ostacolo all'applicazione delle tecniche di analisi testuale, e, al tempo stesso, si delineeranno i principali ambiti applicativi di tali tecniche.

- Nel Capitolo 2 si tracceranno alcune coordinate sulle principali famiglie di analisi del sentiment e sul funzionamento delle principali tecniche di analisi. Si procede con la spiegazione delle metodologie di tipo *machine learning* e con quelle *knowledge-based* fornendo per entrambe alcuni esempi ai fini di rendere più chiara l'esposizione. Per concludere il Capitolo, si espongono alcune tecniche che non rientrano in queste due famiglie, ma rilevanti in letteratura, al fine di produrre una sintesi essenziale ma quanto più completa delle metodologie presenti allo stato dell'arte.

- Nel Capitolo 3 verranno esposti i principali strumenti informatici accessibili attualmente sul mercato. Si vedranno alcune dimostrazioni di *brand monitoring* e si esporranno alcune testimonianze di membri dell'organigramma aziendale di alcune imprese di fama internazionale. Concludendo verrà riportata un'applicazione degli strumenti di *sentiment analysis* in campo politico, portando ad esempio alcuni casi di previsione dell'andamento del gradimento - e del conseguente punteggio in termini di voti - in tema di elezioni.

CAPITOLO 1. CONCETTI INTRODUTTIVI

1.1. Sentiment analysis e opinion mining

L'avvento dei social network e il ruolo che questi stanno assumendo anno dopo anno nel rivoluzionare il concetto di comunicazione, in una società sempre più informatizzata e interconnessa, ha dato modo al mondo delle imprese di ampliare le proprie possibilità e di stringere un rapporto più diretto con il mondo dei consumatori. L'accesso alla rete ha abbattuto i più sostanziali ostacoli in una delle fasi più critiche che costituiscono la relazione azienda/consumatore, ovvero quella dello scambio delle informazioni, rendendo la comunicazione fra le parti veloce e immediata. In particolare, il cosiddetto *eWOM* (*electronic words-of-mouth*), ovvero il passaparola in rete, è fattore fondamentale soprattutto riguardo la valutazione della *brand reputation* o della *customer satisfaction* nei confronti di prodotti o servizi; il monitoraggio dei canali digitali di comunicazione ha un notevole impatto nell'ambito del *brand management*, è indubbiamente utile per valutare le fasi del processo decisionale nell'acquisto di un prodotto ed è eventualmente in grado di influenzarle (Ceron, Curini e Iacus, 2014).

È in questo contesto che prendono piede strumenti come la *sentiment analysis (SA)* e l'*opinion mining (OM)*. Definire cosa si intenda con questi termini, nucleo del presente elaborato, non è cosa da poco dal momento che la letteratura non risponde in maniera univoca nel delineare il loro significato. La rassegna delle loro definizioni, reperite attraverso la fase di ricerca, è particolarmente interessante perché sottende una dinamica dialogica tra i principali autori che riflette l'urgenza del sistema di adeguarsi alla crescita frenetica dei mezzi.

Secondo Ceron, Curini e Iacus (2014) la *sentiment analysis* si pone l'obiettivo di analizzare un sentimento, contenuto ed espresso all'interno di un testo, valutandone non solo la tipologia (positiva o negativa) ma anche l'intensità. Per *opinion mining* invece si intende la tecnica che elabora una ricerca su parole chiave in grado di identificare, per ogni termine, gli attributi (positivo, negativo, neutro) che successivamente permettono di determinare l'opinione associata a ciascuna parola chiave. Gli autori indicano anche la locuzione *opinion analysis* che individua lo studio rivolto alle motivazioni che sono alla base di un *sentiment* positivo o negativo.

Secondo quanto affermato da Farhadloo e Rolland (2016) i termini *sentiment analysis* e *opinion mining* sono utilizzabili come sinonimi e viene identificato come loro primario obiettivo quello di scoprire le opinioni delle persone espresse in un linguaggio scritto; alla parola "sentimento" viene associata un'esperienza personale che porta ad avere una certa opinione in un determinato argomento.

Anche secondo Liu (2012) i due termini sono assimilabili ad uno stesso concetto, ovvero il campo di studio che analizza le opinioni, i sentimenti, le valutazioni, le stime, la predisposizione e le emozioni delle persone verso prodotti, servizi, organizzazioni, individui, tematiche, eventi, argomenti e altri elementi.

Per una maggior fluidità e chiarezza, da qui in avanti si utilizzeranno le accezioni afferenti a quest'ultima definizione che prevede appunto che i due termini siano parte dello stesso concetto.

Molti autori procedono successivamente con una classificazione della *sentiment analysis* in base all'oggetto in analisi.

- *Document-level sentiment analysis*; viene studiato un intero documento (considerato come una singola unità) e viene classificato in base al tipo di opinione che l'intero testo fa trasparire, sia essa negativa o positiva.

- *Sentence-level sentiment analysis*; a questo livello l'analisi si sposta alla singola frase e determina, come nel primo caso, il tipo di opinione che viene espressa. Solitamente ad un'opinione neutrale viene associato un valore nullo. Questo tipo di analisi è strettamente correlata al concetto di *subjectivity classification*, il quale distingue le cosiddette *objective sentences*, quelle frasi che esprimono un dato o un fatto oggettivo, dalle *subjective sentences* che esprimono invece un punto di vista e un'opinione.

- *Entity and Aspect level sentiment analysis*; le prime due classificazioni non riescono a individuare esattamente che cosa sia o meno di gradimento per le persone. Questo tipo di analisi invece, al posto di studiare la costruzione linguistica del testo (documento, paragrafo, frase, periodo) va a osservare l'opinione stessa. È basata sul fatto che un'idea sia composta da un *sentiment* (positivo/negativo) e da un *target* (l'obiettivo conoscitivo). Quando non viene focalizzato il *target* di un'opinione, quest'ultima assume scarsa rilevanza. Ecco dunque che risulta necessario definire gli ambiti e gli aspetti su cui concentrare le analisi testuali (Liu, 2012).

Ceron, Curini e Iacus (2014) contribuiscono, a prescindere dall'oggetto in analisi, a delineare in quattro punti cardine quelli che possono essere definiti i principi che costituiscono le fondamenta di ogni processo di analisi testuale. Non è sufficiente, secondo gli autori, affidarsi incondizionatamente alle capacità computazionali dei calcolatori, i quali, non possedendo le qualità distintive di un essere umano quali l'intelletto e la capacità di ragionamento (in senso proprio), riscontrano gravi difficoltà nell'approcciarsi al mondo della semantica, come anche a quello dell'emotività e del sentimento, tema centrale dell'intero problema. Quindi risulta fondamentale la costante cooperazione fra i software, capaci di raccogliere, sondare, classificare e sintetizzare una mole spaventosa di informazioni, e l'uomo, in grado di comprendere il linguaggio e il significato che esso racchiude.

1. Partendo dal presupposto che ogni modello linguistico quantitativo è sbagliato, ma qualcuno può essere utile, è ragionevole sostenere che procedere con lo studio del *sentiment* di un testo facendo affidamento sulla sola "forza bruta" dei calcolatori, ammassando dosi sempre crescenti di dati, possa risultare controproducente e spesso non porta ad alcun risultato. È dunque necessario fare una distinzione fra *quantità* e *qualità*: non tutte le informazioni sono rilevanti, ed è di fondamentale importanza definire preventivamente obiettivi e benefici ricercati attraverso l'analisi. Inoltre, i computer non sono in grado di cogliere le infinite sfaccettature della lingua scritta e non possono distinguere le sottigliezze in campo semantico, quali ad esempio l'ironia e il sarcasmo. Gli stessi autori, in un'intervista rilasciata al Corriere della Sera (29 gennaio 2014) affermano che *"una macchina non è in grado di comprendere certe espressioni, così come non è capace di associare determinati soprannomi a una figura politica […]. O di interpretare alcune espressioni come 'che bella fregatura'. Ecco perché allora è necessario il lavoro umano."*

2. I metodi quantitativi aiutano l'uomo, non lo sostituiscono. In queste poche parole viene racchiuso un significato importante, ovvero che per riuscire in maniera efficace in un'analisi del *sentiment* sono necessarie cooperazione fra algoritmi automatici e supervisione manuale per opera di esseri umani. Almeno per il momento, le macchine non sono in grado di approcciare una

lingua in tutta la sua complessità come invece riesce all'uomo e per questo motivo le prime non possono lavorare autonomamente senza il secondo.

3. Non esiste una tecnica ideale di analisi testuale; è necessario, a seconda delle esigenze, procedere con la tecnica che è più in grado di essere incisiva e efficace.

4. L'analisi deve essere validata dai dati stessi; questa si distingue fra tecniche supervisionate, ovvero tecniche in cui le categorie semantiche sono note a priori, e non supervisionate, dove invece sono determinate a posteriori. Tale tipo di validazione può essere particolarmente gravosa perché richiede la classificazione (fatta dall'uomo) dei testi nelle varie categorie semantiche attraverso un incrocio di dizionari di termini e vocaboli (come verrà definito nel Capitolo 2).

1.2. Il contesto di applicazione: le reti sociali.

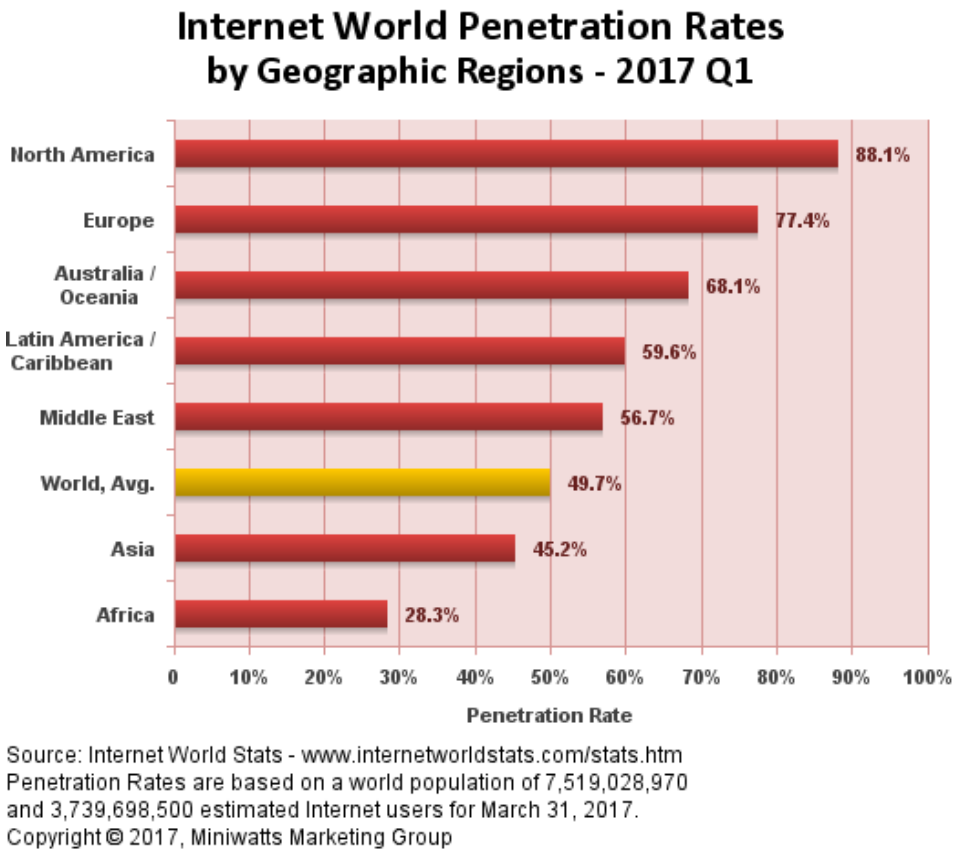
Risulta particolarmente utile, in via introduttiva, mettere chiarezza sul contesto all'interno del quale l'analisi del *sentiment* si trova ad operare. La dicitura *social network* è di frequente utilizzo, ciò nonostante spesso si rischia di confonderla con il termine *social media*, che ha una valenza differente. La prima si riferisce, come definito da Scott (2000) a una qualunque struttura, formale o informale, che comprende un insieme di persone o organizzazioni, assieme alle loro rispettive relazioni; come si può notare, non vi è alcun riferimento a Internet, in quanto il concetto di *social network* comprende un insieme più ampio dei *social media*.

Il secondo invece può essere definito come un gruppo di applicazioni Internet basate sui presupposti ideologici e tecnologici del *Web 2.0*, che consentono la creazione e lo scambio di contenuti generati dagli utenti (Wikipedia, 2017). Si può affermare che i secondi siano un sottoinsieme dei primi, in termini teorici; è attraverso i *social media* che operano prevalentemente gli strumenti di *sentiment analysis*, ragion per cui, una volta espresse le dovute precisazioni terminologiche, nel resto dell'elaborato il discorso prenderà in maggior considerazione quest'ultima categoria.

Ciò che si vuole cercare di comprendere è, però, per quale motivo i *social media* siano il contesto più utilizzato per l'analisi testuale.

Una prima ragione è di facile individuazione: come già affermato in precedenza, i *social media* sono un'inesauribile fonte informativa aggiornata in tempo reale, nonché la più ampia fra quelle fruibili dalle imprese. A marzo 2017 risulta che gli utilizzatori di Internet nel mondo sono circa 3,7 miliardi, che equivale ad un tasso di penetrazione (ovvero la percentuale di persone considerate utenza di Internet in rapporto alla popolazione totale) del 49,7%: quasi una persona su due, al mondo, ha la possibilità di accedere ad Internet. In vetta alla classifica dei maggiori utilizzatori della Rete troviamo l'Asia, con 1,8 miliardi di utenti; l'America del Nord invece primeggia in termini di tasso di penetrazione attestandosi sul 88,1% (Internet World Stats, 2017). I fatti esprimono, in maniera lampante, una tendenza ad una crescita sostenuta di questi dati: basti confrontare i risultati aggiornati a luglio 2013 (ottenuti dalla stessa fonte sitografica) che sono riportati in Ceron, Curini e Iacus (2014) con i dati sopra elencati; l'utenza mondiale di Internet risultava essere di 2,4 miliardi (ben 1,3 miliardi in meno del 2017) e il tasso di penetrazione era del 34,3% (15,4 punti percentuali in meno del 2017).

Figura 1: Tasso di penetrazione di Internet nel mondo.



Alla luce di questi dati si può comprendere il motivo per cui le aziende stiano via via concentrando sempre di più i propri sforzi (soprattutto a livello economico) nelle strategie di *social media management* e nelle tecniche di *sentiment analysis*; un gran numero di utenti, ogni giorno, inonda il web di opinioni, pensieri e consigli che se intercettati in maniera intelligente

dalle imprese costituiscono un vero e proprio tesoro potenzialmente capace di fornire dati utili e funzionali da utilizzare come base per le decisioni presenti e future.

Un'altra ragione che porta a un utilizzo dell'analisi testuale nel contesto dei *social media* è dovuto alla qualità dei dati raccolti, oltre che alla quantità. La ricerca di feedback attraverso la *sentiment analysis* si avvicina molto a tecniche di ricerca di marketing quali il *focus group* e i sondaggi, nonostante queste ultime siano due metodologie molto differenti fra loro; la *sentiment analysis* riesce infatti a conciliare le risposte personali e soggettive di un *focus group* e l'ampiezza di feedback raggiungibile attraverso un sondaggio ma con una minor onerosità sia sul piano di progettazione sia sul piano economico giungendo ad una maggior efficacia con uno sforzo minore.

Ulteriore motivazione di sviluppo attraverso i *social media* delle tecniche di analisi testuale risiede nel duplice metodo di approccio che può essere adoperato dalle imprese in questo contesto, proprio a causa della natura dei *social media* stessi.

Un primo approccio è quello cosiddetto *top-down*: le imprese creano delle *community* su prodotti o brand, condividono le informazioni su novità e fatti rilevanti, possono avere un contatto diretto e tempestivo con i consumatori, tutti fattori che permettono di consolidare e trasmettere valori di brand o aziendali (Zarella, 2009). Questo tipo di approccio è mirato al *micromarketing*; le imprese puntano all'utilizzo di risorse di *data mining* per modificare prodotti e programmi di marketing per assecondare le preferenze dei singoli individui e gruppi locali di clienti specifici, adottando a tutti gli effetti una forma "estesa" di *marketing individuale* (Kotler, Armstrong, Ancarani e Costabile, 2015).

Il secondo approccio è viceversa quello definito come *bottom-up*: viene considerato il *social media* come una moderna agorà da studiare in modo appropriato e da cui estrarre informazioni utili per fornire un aiuto prezioso per comprendere l'evoluzione di fenomeni sociali complessi (Ceron, Curini e Iacus, 2014). Sulla base di quest'ultimo pensiero si sono creati i presupposti per quello che viene definito *nowcasting*, ovvero *previsioni sul presente*, identificando in tempo reale le dinamiche che si sviluppano a livello sociale fra gli utenti.

Concludendo, alla luce di quanto analizzato finora, risulta evidente l'importanza dei *social media* come contesto all'interno del quale gli strumenti di *sentiment analysis* si trovano ad operare.

La crescente affluenza di utenti al mondo del web e la loro conseguente maggior partecipazione in maniera attiva fa sì che le imprese focalizzino la loro attenzione su questo ambito, evolvendo la loro visione di ricerca di marketing.

1.3. Sfide e ambiti applicativi.

Le tecniche di *sentiment analysis* e *opinion mining*, nonostante la loro inequivocabile utilità sotto diversi punti di vista, non possono tuttavia fare a meno di scontrarsi con alcune problematiche, sia da un punto di vista teorico che da quello applicativo.

Una prima sfida particolarmente rilevante è data dalla difficoltà da parte di piccole e medie imprese di equipaggiarsi dei mezzi informatici in grado di svolgere processi complessi come quelli richiesti dalle analisi testuali o da tecniche di brand management, soprattutto a causa del loro ingente costo. Per contrastare questa difficoltà, sono nati alcuni software (per citarne solo alcuni: *Brandwatch*, *Lithium*, *Mantra*) i quali dispongono di *Cloud Computing technologhy*, che permette la condivisione delle risorse computazionali senza obbligare le imprese a dover acquistare l'intera infrastruttura (Benedetto e Tedeschi, 2016). Con l'acquisto del software di monitoraggio le imprese ottengono anche la possibilità di usufruire dell'intero patrimonio informatico messo a disposizione dalle aziende specializzate nel campo del *brand management* e della *SAOM (sentiment analysis and opinion mining)*.

È altresì vero che, a causa del loro bacino di utenza più ampio e delle loro capacità di investimento più elevate, sono le aziende di dimensioni maggiori le prime a beneficiare degli strumenti di analisi testuale. Essendo necessario l'inserimento all'interno dell'organigramma aziendale di un'unità predisposta all'attività di *social media management*, si dimostra essere ancora più complicato l'accesso da parte delle piccole e medie imprese a questo tipo di tecnologia.

Spostando l'attenzione sull'ambito tecnico, Farhadloo e Rolland (2016) si adoperano nel fornire una serie di *challenges* al processo di analisi testuale che si frappongono fra gli obiettivi che ci si pone e il loro effettivo raggiungimento. Come abbiamo già definito in precedenza, il linguaggio umano è un insieme complesso di termini e concetti che a seconda di come vengono posti in relazione fra loro portano ad un'interpretazione specifica di una data affermazione, ed è questo il più grande ostacolo interpretativo da parte di software e algoritmi. In particolare si possono elencare alcune difficoltà nell'estrapolazione di sentimenti e opinioni da parte dei computer proprio legate alla complessità del linguaggio:

- *Sinonimi e polisemia*: nel primo caso viene descritta la stessa informazione utilizzando termini diversi, nel secondo invece vengono utilizzati termini identici per riferirsi a concetti differenti; in contesti differenti, o se usate da diverse persone, le stesse parole

assumono significati diversi e questo può creare difficoltà nella creazione di un metodo di codificazione che possa essere corretto e veritiero.

- *Sarcasmo*: capire una frase espressa con sarcasmo richiede una profonda comprensione del contesto all'interno della quale si trova, dell'argomento, del linguaggio e delle persone che sono coinvolte. Avere accesso alla totalità di tali informazioni è un procedimento complicato già di per sé; richiedere queste competenze ad un computer si rivela essere particolarmente complesso.

- *Frasi composte*: sono due frasi indipendenti che sono collegate da congiunzioni come "e", "o", "ma", "per". Le frasi "I bambini si sono divertiti in spiaggia ma noi no" oppure "Il servizio è stato impeccabile ma non posso affermare che questo sia il miglior ristorante della città" sono dei chiari esempi di questo tipo di problema. Nella frase vengono esplicitate due opinioni contrastanti, una con *sentiment* positivo e una invece con *sentiment* negativo; questo può costituire un problema nel momento della codifica automatica da parte di un computer.

- *Dati non strutturati*: i feedback che vengono presi in analisi solitamente assumono la forma di "testi grezzi", ovvero nella loro forma più basilare; un passaggio molto complicato e laborioso consiste nel trasformare questi dati grezzi in dati semi-strutturati, associando i testi a *tag* e ad altri *markers* che aiutino i calcolatori a separare i contenuti semantici gli uni dagli altri. Tale procedimento spesso viene abbinato ad un lavoro manuale che richiede tempo e dispendio di risorse, risultando un'ulteriore sfida al corretto svolgimento dell'analisi.

Concludendo, gli autori affermano che la cosiddetta *Computational Intelligence*, ovvero quella branca degli studi sull'Intelligenza artificiale che si concentra sull'apprendimento, l'adattamento e l'evoluzione di quei programmi che si possono definire, in un certo senso, intelligenti (Benedetto e Tedeschi, 2016), svolge un ruolo fondamentale nei confronti della *sentiment analysis* e si è dimostrata un mezzo potente per comprendere le percezioni dei consumatori in relazione a prodotti o servizi.

Nonostante ci siano stati notevoli avanzamenti durate la breve storia di questo campo di studi, c'è ancora una gran quantità di lavoro da fare. La maggior parte della discussione finora è stata indirizzata nel decifrare i contenuti semantici dei testi scritti, ma questa ricerca ha

cozzato contro alcuni scogli linguistici notevoli. Ciononostante, si è potuto notare come la ricerca ha proposto metodi che scovano sentimenti e opinioni e che corrispondono in maniera significativa con i dati ottenuti attraverso analisi della *customer satisfaction*. Ciò che rimane nebuloso è il verificare se questi metodi siano o meno generalizzabili a tutti i contesti e se le tecniche probabilistiche di *computational intelligence* possano essere effettivamente versatili. Le opportunità di uno sviluppo nella ricerca sono tuttavia ampie; questo campo di studi porterà un cambiamento radicale nella comprensione dei consumatori da parte delle organizzazioni e probabilmente in come questi ultimi percepiscono e valutano prodotti e servizi (Farhadloo e Rolland, 2016).

A conclusione di questo Capitolo si riporta una breve spiegazione dei principali ambiti applicativi di questo campo di studi e delle tecniche di analisi testuale ad esso correlati.

Ceron, Curini e Iacus (2014) presentano una sintesi delle macroaree che sono state prevalentemente coinvolte nell'analisi del *sentiment* negli ultimi anni. Tale esposizione (*Tabella 1*) non punta in alcun modo a presentare un elenco definitivo e puntuale di ogni singolo caso di studio, piuttosto vuole fornire al lettore una panoramica quanto più varia e completa possibile delle potenzialità delle tecniche di analisi testuale e di come sono già state applicate per effettuare *nowcasting (previsione in tempo reale)* e *forecasting (previsioni sul lungo termine).*

Tabella 1: Argomenti studiati sui social media in relazione al tema delle "previsioni"

AREA	PREVISIONE/STIMA	FONTE	CITAZIONI
ECONOMIA	Indici in borsa	Twitter; Google; Blog	Bollen *et al.*, 2011; Gilbert e Karahalios, 2010; Preis *et al.*, 2012; Zhang e Fuehres, 2011; Zhang *et al.*, 2012
	Volatilità dei mercati finanziari	Forum	Antweiler e Frank, 2004
	Indicatori macroeconomici	Google	McLaren e Shanbhogue, 2011
EPIDEMIOLOGIA	Diffusione influenza e altre malattie	Google; Twitter	Achrekar *et al.*, 2013; Cook *et al.*, 2011; Freifeld *et al.*, 2008; Ginsberg *et al.*, 2009; Lampos e Cristianini 2012; Signorini *et al.*, 2011; Valdivia *et al.*, 2010;

	Probabilità di malattie e decessi	Wikipedia e varie	Radinsky e Horvitz, 2012
MARKETING	Acquisto/consumo di prodotti	Blog; Google	Gruhl *et al.*, 2005; Liviu, 2011; McLaren e Shanbhogue, 2011
	Incassi al box office	Twitter	Asur e Huberman, 2011
POLITICA	Risultati elettorali	Twitter (principalmente); Facebook	(Ceron, Curini e Iacus, 2014)
	Popolarità dei politici	Twitter	(Ceron, Curini e Iacus, 2014)
	Rivolte	Google; Wikipedia e varie, Twitter	Kalev, 2011; Koehler-Derrick e Goldstein, 2011; Radinsky e Horvitz, 2012
PSICOLOGIA	Umore e stati d'animo	Twitter	Lansdall- Welfare, 2012
	Felicità	Twitter	(Ceron, Curini e Iacus, 2014)
SISMOLOGIA	Individuazione di terremoti	Twitter	Sakaki *et al.*, 2013
SOCIETÀ	Vincitori di concorsi televisivi	Twitter	Ciulla *et al.*, 2012
	Auditel	Twitter	–
	Risultati sportivi	Twitter	UzZaman *et al.*, 2012
	Vincitori Oscar	Varie	Bothos *et al.*, 2010; Liviu, 2011

Fonte: Ceron, Curini e Iacus (2014), p. 14

Come si può evincere osservando la *Tabella 1*, la *sentiment analysis* è uno strumento in grado di adattarsi ad un gran numero di situazioni differenti. Le molteplici produzioni testuali incanalate all'interno dei *social media* rappresentano le preoccupazioni, le intenzioni e le propensioni che le persone hanno nei confronti degli argomenti a cui più tengono. L'utilizzo degli strumenti di *sentiment analysis* dipende dunque dagli enti che ne fruiscono; non sono solo un mezzo a disposizione delle aziende per verificare l'efficacia delle proprie scelte di marketing ma anche un mezzo a disposizione di enti nazionali o governativi per studiare le preferenze politiche più affermate, l'umore, lo stato d'animo, o addirittura il propagarsi di un terremoto o la diffusione di determinate malattie. La *sentiment analysis* si rivela essere un potente strumento che, a seconda degli utilizzatori e degli argomenti trattati, è in grado di sondare un ampio quantitativo di informazioni in tempo reale e in certi casi di prevedere l'andamento di alcuni fenomeni nel breve periodo.

Nel Capitolo seguente verranno esposte le principali tecniche di *sentiment analysis*, delineandone il funzionamento generale e le principali differenze.

CAPITOLO 2. STRUMENTI E METODI DI ANALISI

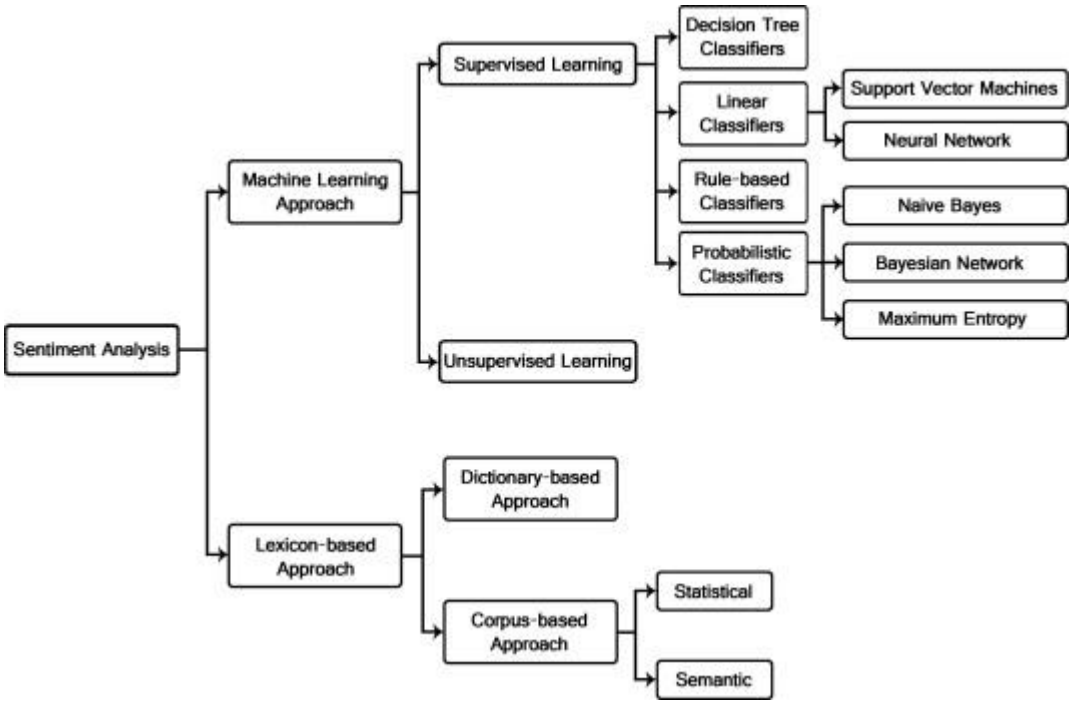
2.1. Convertire opinioni in dati: i principali strumenti

La letteratura in materia di *sentiment analysis* fornisce una corposa lista di categorie e suddivisioni per quanto riguarda le tecniche di applicazione pratica, a seconda degli autori considerati. Nonostante la varietà di sottocategorie proposte dai testi sul tema, Benedetto e Tedeschi (2016) sintetizzano l'argomento raggruppando le tecniche di analisi testuale in due macrogruppi i quali seguono due approcci differenti:

- il *machine learning approach*, ovvero un approccio basato sull'apprendimento automatico il quale si articola a sua volta in *supervised* e *unsupervised*;

- le *knowledge-based techniques* (o anche *lexicon-based approach*) ovvero le tecniche basate sul lessico.

Figura 2: Tipologie di sentiment analysis



Fonte: Medhat, Hassan e Korashy (2014), p. 1095

Come si può notare nel diagramma esposto nella *Figura 2*, costruito da Medhat, Hassan e Korashy (2014), esistono un gran numero di sottoclassi contenute nelle due macroaree prese in considerazione. In questo lavoro si considera un'analisi dell'argomento condotta ad un livello quanto più ampio possibile, ponendo particolare attenzione sulle caratteristiche peculiari delle due macroaree e accennando solo velocemente alle suddivisioni particolari contenute al loro interno. Per completezza viene dunque riportato in *Figura 2* l'intero schema che rappresenta in maniera più dettagliata possibile, almeno secondo la gran parte della letteratura su questo tema, le suddivisioni in sottogruppi delle tecniche di analisi testuale del *sentiment*. Alla fine del Capitolo si farà riferimento ad altre metodologie che esulano dalla rappresentazione fornita dal diagramma in quanto non ricomprese né nell'una né nell'altra categoria ma che comunque assumono una certa rilevanza nel contesto della *sentiment analysis.*

Prima però di affrontare l'argomento, è necessario fare un passo indietro e rispondere alla seguente domanda: come è possibile trasformare dei testi in dati statisticamente rilevanti e in grado di fornire indicazioni riguardo il *sentiment* di chi li ha prodotti?

2.1.1. Il preprocessing

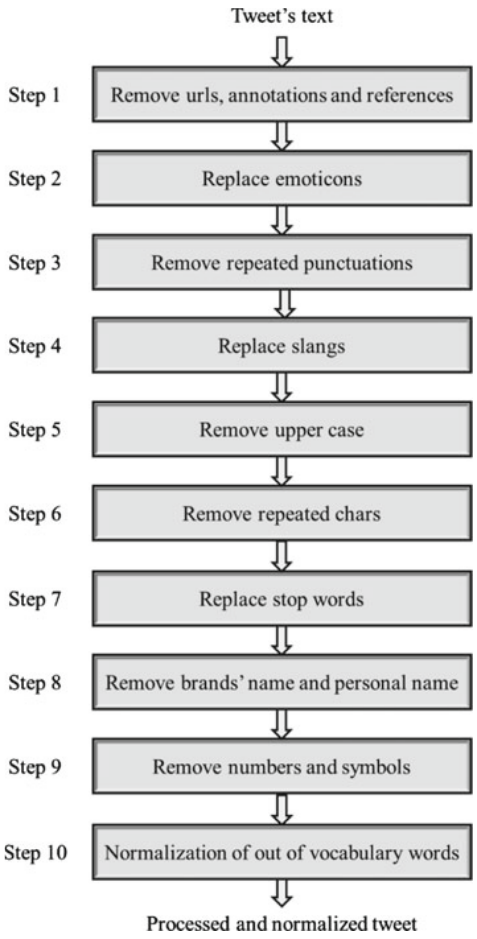
Attraverso la fase di *preprocessing* un testo viene trasformato in modo tale che un algoritmo sia successivamente in grado di trattarlo. È un processo manuale, che dev'essere necessariamente compiuto dall'uomo per far sì che il lavoro dei software sia facilitato e possa raggiungere i risultati prefissati, ovvero estrapolare dal documento la relativa opinione espressa dallo scrivente. Un testo solitamente fa parte di un insieme più ampio di documenti, definito *corpus*; una collezione di *corpus* viene chiamata *corpora*. A seconda dei metodi utilizzati e delle finalità degli utilizzatori, si può focalizzare l'analisi a più gradi di profondità e di dettaglio e quindi porre maggior attenzione sul testo preso singolarmente o sull'intero insieme di *corpus*.

Una prima fase del *preprocessing* consiste nell'alleggerire il carico informativo da analizzare eliminando l'informazione relativa all'ordine con cui le parole figurano all'interno del testo. Si definisce il risultato di tale operazione come *bag of words* (letteralmente "contenitore di parole") in quanto le parole non assumono più un significato in base a come sono poste in relazione fra loro, ma viene considerato piuttosto il significato della singola parola presa singolarmente.

Dato questo come obiettivo, Benedetto e Tedeschi (2016) propongono uno schema da seguire per riuscire a raggiungere una *bag of words* alleggerita da ogni informazione superflua a partire da un testo completo, rendendo chiari i procedimenti attraverso una serie di step

(*Figura 3*). Essi partono dall'operazione di selezione di un *tweet*, passano attraverso il procedimento di *preprocessing* e terminano con un testo "processato" e "normalizzato".

Figura 3: Diagramma del procedimento di preprocessing



Fonte: Benedetto e Tedeschi (2016), p. 361

Gli autori, con l'intento di rendere esplicito il procedimento di lavoro, propongono un esempio pratico partendo da un *tweet* reso anonimo e privato di riferimenti riguardo le aziende *competitor* prese in causa. Tale testo, in sintesi, focalizzando l'attenzione sugli *step 1* e *10* evolve in maniera sostanziale.

Nello *step 1* si trova il seguente testo:

RT @AUTHOR_MENTION Omg, james i h8 you wth you talkin about >:(,cuz it's so coooool!!1! I can imagine him, taking it with meeeee,would be so awesome adding his messages to fav. .. #BRAND_lover btw the problem is my phone is COMPETITOR_NAME... not BRAND_NAME :(

Come si può notare, un testo di questo tipo nella sua forma originale (e quindi più grezza) può risultare di difficile interpretazione da parte di un algoritmo; se il significato e il *sentiment* possono essere compresi, con la dovuta attenzione e non nella totalità dei casi in maniera chiara e lampante, da una persona che si trovi a leggere il *tweet*, è molto più complicato che un software sia in grado di rilevarne il significato.

Una volta conclusa la trasformazione del testo attraverso il *preprocessing* e raggiunto quindi lo *step 10* del procedimento, il tweet assume una forma completamente diversa:

god hate hell talking about angry cool brilliant can imagine following be awesome adding messages favourites by the way the problem phone not sad

Come si può facilmente notare, giunti al fondo della lista di *step* da percorrere il testo prende una forma più chiara e "intellegibile" da un software di analisi; ad esempio, forme di espressione dello stato emotivo come *smile* oppure termini ridotti, slang regionali (o giovanili), o, ancora, punteggiatura e storpiature delle parole vengono esplicitati e semplificati, portando ad un testo conclusivo ordinato ma soprattutto convertibile in un pacchetto di dati utile per l'analisi.

Questo tipo di testo rappresenta un esempio di *bag of words* in grado di fungere da base operativa per iniziare un'analisi testuale. Tale procedimento però, come già anticipato, dovendo essere compiuto manualmente testo per testo risulta particolarmente gravoso sia in termini di tempo che di risorse umane impiegate; per condurre una *sentiment analysis* efficace possono essere necessari migliaia di testi e di conseguenza una mole non indifferente di lavoro in fase di *preprocessing*.

Il tema del *preprocessing* è stato ampiamente trattato dagli analisti del settore in quanto punto di partenza per ogni approccio di analisi testuale e dunque pilastro fondamentale da cui muovere l'intero processo di *sentiment analysis*. Ne dà conto in modo approfondito la letteratura e in particolare quanto pubblicato da Benedetto e Tedeschi (2016) e da Haddi, Liu e Shi (2013) sull'approfondimento tecnico del funzionamento delle fasi del *preprocessing*.

2.1.2. Lo stemming

Una volta completata la prima fase di *preprocessing*, e avendo dunque ottenuto un testo libero da informazioni irrilevanti presentato in maniera ordinata, si procede con una seconda fase definita *stemming*. Lo *stemming* è un processo che riduce una parola al suo *stem* (stilema),

ovvero la sua radice. La radice può presentarsi sotto varie forme, a seconda delle necessità e delle scelte prese precedentemente l'analisi: può comparire come un termine di senso compiuto (ad esempio *house, man, product, happy*), o al contrario, può non presentarsi sotto la forma di una parola dal senso compiuto di per sé, ma può essere a sua volta utilizzata per generare parole aggiungendo dei suffissi. Per esempio: le parole *fish, fishes,* e *fishing* sono riconducibili allo stesso *stem, fish*, che è a sua volta una parola dal senso compiuto; al contrario le parole *study, studies* e *studying* vengono ricondotte a *studi*, che in inglese non ha un significato proprio. I motori di ricerca applicano lo *stemming*, tradizionalmente, per migliorare la possibilità di ottenere come risultato della ricerca forme differenti della stessa parola, trattandole alla stregua di sinonimi dato che concettualmente parlando appartengono alla stessa famiglia di termini (Bonzanini, 2015).

Ceron, Curini e Iacus (2014), con l'intento di rendere più facilmente comprensibile questo passaggio del procedimento, hanno costruito un esempio pratico di applicazione dello *stemming* il quale viene di seguito riportato (anch'esso in forma ridotta, attraverso l'esposizione dei passaggi chiave). I testi su cui si vuole applicare lo *stemming* sono, al contrario di quanto utilizzato finora nella conduzione del discorso, in lingua italiana, ma i ragionamenti fin qui spiegati possono essere applicati a qualunque lingua indistintamente.

- Testo 1: *il* ***nucleare*** *conviene perché è* ***economico***.
- Testo 2: *il* ***nucleare*** *produce* ***scorie***.
- Testo 3: *il* ***nucleare*** *mi fa* ***paura*** *per le* ***radiazioni***, *le* ***scorie*** *e non riduce l'****inquinamento***.

Si supponga, per semplicità, che il procedimento di *stemming* abbia evidenziato i termini in grassetto come parole rilevanti ai fini dell'analisi. Si procede delineando una matrice in cui ogni riga rappresenta un testo e ogni colonna rappresenta uno *stem*; avremo dunque gli stilemi *s1 = nucleare, s2 = paura, s3 = radiazioni, s4 = inquinamento, s5 = scorie, s6 = economico* e via dicendo. Viene analizzato ogni testo verificando la presenza o meno di ogni stilema preso in considerazione, associando ad 1 la presenza dello *stem* e a 0 la sua assenza. Per fare qualche esempio, prendendo in considerazione il Testo 1 il vettore di *stem* assume una forma del tipo S_1 = (*s1, s2, s3, s4, s5, s6*) = (1, 0, 0, 0, 0, 1), mentre il Testo 2 S_2 = (1, 0, 0, 0, 1, 0). Ogni testo viene ricompreso in una categoria semantica D_k, k = 1, …, K, dove K è il numero totale di categorie semantiche. Ponendo ad esempio K = 2 si potrebbero ipotizzare come categorie semantiche le opinioni D_1 = *a favore* e D_2 = *contro* un determinato argomento. Gli autori

procedono successivamente con il delineare una tabella che riassume le informazioni esposte fino a questo punto (*Tabella 2*).

Tabella 2: Esempio di matrice di stemming

Post	D_i	s1 nucleare	s2 paura	s3 radiazioni	s4 inquinamento	s5 scorie	s6 economico	…
testo 1	a favore	1	0	0	0	0	1	…
testo 2	N/A	1	0	0	0	1	0	…
testo 3	contro	1	1	1	1	1	0	…
testo 4	contro	1	1	1	1	1	0	…
testo 5	a favore	1	0	1	1	1	0	…
…	…	…	…	…	…	…	…	…
testo n	a favore		0	1	0	0	1	…

Fonte: Ceron, Curini e Iacus (2014), p. 33

Come evidenziato dalla tabella, attraverso lo *stemming* si raggiunge l'obiettivo di trasformare dei testi "grezzi", privi cioè di un'effettiva capacità di comunicare il proprio contenuto, agli algoritmi che dovranno poi analizzarlo, in dati facilmente riconoscibili dagli strumenti di analisi testuale. Il principale problema che si può riscontrare in questa fase del procedimento è indubbiamente la mole di informazioni che è necessario vagliare; a seconda dei casi, gli *stem* possono raggiungere migliaia di termini, e verificare la presenza o meno di tali termini nei testi può risultare un'operazione di difficile attuazione.

2.2. Machine learning Approach

La prima famiglia di analisi del *sentiment* prende il nome di *machine learning approach*, ovvero un approccio basato sull'apprendimento automatico. Questo metodo sfrutta l'utilizzo di algoritmi per condurre una *sentiment analysis*; la famiglia degli approcci che hanno fondamento sull'apprendimento automatico può essere ulteriormente suddivisa in due tipologie: *supervised* e *unsupervised*.

2.2.1. Supervised method

La prima, più comunemente usata, si pone come obiettivo quello di far apprendere al computer un sistema di classificazione che è stato precedentemente progettato, costruito su misura per il singolo caso di analisi che si vuole affrontare (Benedetto e Tedeschi, 2016). In sostanza, ciò si può ottenere attraverso la costruzione di alcuni *training set* affinché il computer sia in grado di comprendere un certo input e di fornire il relativo output sotto forma di classificazione del testo rispetto un determinato orientamento.

Il *training set* viene creato manualmente per fare sì che ad ogni input venga correlato il corretto output, definendo così la funzione di apprendimento *f*. L'idea alla base di questo approccio è che attraverso il *training set* il computer sia in grado di apprendere la correlazione fra input e output e sintetizzarla in una funzione *j* che sia un'approssimazione di *f*. Se l'approssimazione risulta accettabile allora il sistema dovrebbe essere in grado di fornire risultati simili a quelli ottenuti con l'ausilio del *training set*.

I punti critici dell'intero sistema sono la creazione del *training set* e la numerosità dei testi presi in esame dallo stesso; intuitivamente, se il *training set* dispone di un numero ridotto di dati le probabilità che questo fornisca l'output corretto è altrettanto ridotta; viceversa se il *training set* è composto da un numero ragguardevole di dati allora è più probabile che l'output che quest'ultimo si troverà a elaborare sia quello corretto.

Benedetto e Tedeschi (2016) provvedono a delineare i punti cardine da seguire in un procedimento di analisi del *sentiment* nel caso del metodo *machine learning* nella forma *supervised*.

1. Determinare il tipo di testi che compongono il *training set* classificandoli e creandoli manualmente.

2. Decidere la rappresentazione degli input della funzione *f* basandosi sulla forma assunta dagli input stessi e su come sono rappresentati.

3. Strutturare la forma della funzione di apprendimento *f* e degli algoritmi da utilizzare.

4. Procedere con il *training* dell'algoritmo.

5. Valutare l'accuratezza della funzione risultante dall'elaborazione del sistema.

La letteratura propone una vasta gamma di tecniche di analisi *machine learning* basate su questa procedura.

Le più utilizzate sono la tecnica *Naïve Bayes* e quella *Support Vector Machine (SVM)* (Benedetto e Tedeschi, 2016).

Anche Ceron, Curini e Iacus (2014) delineano un sistema di analisi *supervised* definito come *integrated Sentiment Analysis (iSA)* il quale pone il focus sul vantaggio di considerare un'analisi aggregata delle opinioni rispetto ad un'analisi individuale.

2.2.2. Unsupervised method

Questo secondo metodo di utilizzo delle tecniche *machine learning* è decisamente meno comune nella pratica e al contempo meno presente nei testi riguardanti l'analisi del *sentiment*.

Nell'approccio *unsupervised* infatti è più complesso ottenere una stima che si avvicina al dato corretto in quanto, sebbene sia presente un *training set*, non si è in presenza di una classificazione degli output; per l'analisi si dispone unicamente degli input, ma non si conoscono né gli output né tantomeno la correlazione fra questi e gli input (Brownlee, 2016).

Questa tipologia di analisi non è mirata all'ottenimento di una relazione fra i dati immessi e una determinata classificazione dello stesso in quanto non dispone dei mezzi necessari per poter conseguire un tale obiettivo; il focus viene principalmente indirizzato verso il raggruppamento dei testi in insiemi semantici coerenti e nello studio della composizione nel *corpus* preso in analisi.

Tra le tecniche *unsupervised* troviamo tecniche di *data mining* o *text mining* tra cui figura la cosiddetta *cluster analysis*. Per *data mining* si intende quel filone di tecniche volte alla ricerca di una regolarità nei dati, e conseguentemente per *text mining* si intende quell'insieme di tecniche in grado di riscontrare una regolarità all'interno dei testi (Ceron, Curini e Iacus, 2014).

La tecnica definita come *cluster analysis* è la più diffusa nell'ambito delle metodologie *machine learning unsupervised*. Questa ha fondamento nella possibilità di definire una distanza (semanticamente parlando) fra gli oggetti che si ha intenzione di classificare e di definire dei

raggruppamenti quanto più possibile omogenei tra loro basandosi su tale distanza predefinita. Una volta però ottenuti i gruppi è necessario andare ad osservare al loro interno per verificare in cosa gli elementi siano simili in termini di argomenti trattati, e per quale motivo differiscano da altri gruppi, come approfondito e analizzato nei lavori di Agarwal e Mittal (2016) e di Ceron, Curini e Iacus (2014).

2.3. Knowledge-based techniques

La seconda classe di tecniche per la *sentiment analysis* viene definita come *knowledge-based*, o, come viene chiamata in alcuni testi, *lexicon-based* (Benedetto e Tedeschi, 2016). Il fulcro del funzionamento di questo tipo di tecniche sussiste nello sfruttamento di alcune risorse lessicali, come ad esempio i *dizionari ontologici*, per catalogare le opinioni contenute all'interno dei documenti.

I *dizionari ontologici* sono dei particolari dizionari i quali associano ad ogni parola un relativo peso in termine di polarità, positiva o negativa. A seconda della loro polarità, i termini vengono contrassegnati con un peso negativo o positivo che, sommato alla polarità di tutti i termini contenuti all'interno del documento, fornisce l'orientamento generale del testo preso per intero. Solitamente i *dizionari ontologici* suddividono i termini in una scala che va da -1 (parola con significato estremamente negativo) a 1 (parola con significato estremamente positivo) passando per 0 (parola con significato neutrale).

Si può definire, per ricavare il *sentiment* dell'intero testo, la seguente relazione (Benedetto e Tedeschi, 2016):

$$sentiment_{score} = \sum_{i=0}^{n} sentiment\ (word_i)$$

Nella relazione $sentiment_{score}$ rappresenta il punteggio complessivo del testo, *n* corrisponde al numero totale di termini contenuti nel testo e $sentiment\ (word_i)$ rappresenta il valore assegnato al termine i-esimo contenuto nel documento.

Al momento il *dizionario ontologico* più utilizzato, nonché il più voluminoso, è *SentiWordNet*, basato sul dizionario *WordNet*.

WordNet è un ampio database lessicale della lingua inglese, costruito nel 1985 da alcuni linguisti e psicologi dell'Università di Princeton; successivamente è stato tradotto in molte lingue fra cui anche l'italiano. Esso è costruito attraverso dei *synset (synonymus sets)* i quali

sono degli insiemi di sinonimi che esprimono un determinato concetto; ognuno di questi *synset* viene messo in relazione con degli altri insiemi di sinonimi a seconda del collegamento concettuale che sussiste fra loro.

In Agarwal e Mittal (2016) si può ritrovare una lista delle tipologie di collegamenti che è possibile riscontrare all'interno del dizionario *WordNet* (*Tabella 3*):

Tabella 3: Alcune tipologie di collegamenti fra synset in WordNet

Relazione	**Descrizione**
Iperonimia	Collega un *synset* con uno più specifico. Per esempio, "letto" concettualmente viene collegato con un *synset* più specifico come "mobile".
Iponimia	Questa relazione è transitiva: se una poltrona è un tipo di sedia, e una sedia è un tipo di mobile, allora la poltrona è un tipo di mobile.
Meronimia	Y è meronimo di X se Y è una parte di X ("finestra" è meronimo di "edificio").
Antinomia	Relazione fra termini con significato opposto.
Troponimia	Il verbo Y è troponimo del verbo X se l'attività Y comprende l'attività X in qualche maniera ("balbettare" e troponimo di "parlare").

Fonte: Agarwal e Mittal (2016), p. 64

SentiWordNet è un'applicazione del metodo *knowledge-based* sviluppata da Esuli e Sebastiani (2006) al fine di costruire un fondamento lessicale, basato su *WordNet*, in grado di fungere da dato per una *sentiment analysis lexicon based*.

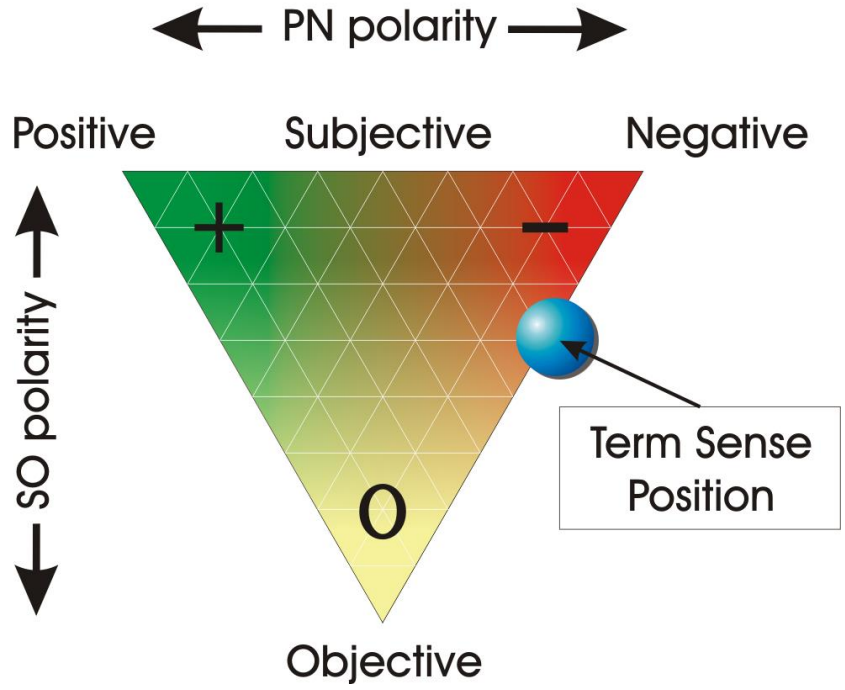
La versione più aggiornata è *SentiWordNet 3.0* ed è disponibile gratuitamente al pubblico. Questo tipo di software lavora in maniera differente da quanto sopra riportato. Il tipo di classificazione dei termini non è effettuato in base ad un solo asse di polarità (ai cui estremi troviamo positivo e negativo) bensì è fatto considerando una duplice direzione della polarità:

- Positivo – Negativo
- Soggettivo – Oggettivo

Vengono assegnati in totale tre punteggi differenti ad ogni singolo termine: un punteggio da 0 a 1 per il valore *negatività*, un punteggio da 0 a 1 per il valore *positività* e un punteggio da 0

a 1 per il valore *soggettività*. Questi valori, messi assieme, forniscono il posizionamento complessivo e la polarità del termine in questione.

Figura 4: Rappresentazione grafica della classificazione di un termine in SentiWordNet



Fonte:(http://ontotext.fbk.eu/sentiwn.html)

La *Figura 4* propone una rappresentazione grafica di quanto affermato finora riguardo la classificazione in base alla polarità di un termine.

Nell'asse che si muove in orizzontale (PN polarity) le parole si collocano in relazione alla loro positività o negatività; nell'asse che si muove in verticale (SO polarity) le parole si collocano in base al loro grado di oggettività o soggettività. Una parola, ad esempio, che si collochi in prossimità del punto colorato in blu avrà una polarità PN relativamente alta e sbilanciata verso un valore negativo, mentre la sua polarità SO è perlopiù neutrale. A seconda dei termini analizzati ci si sposta all'interno del triangolo osservando di volta in volta il posizionamento del termine in esame.

Per una maggior chiarezza, si propone il seguente esempio. L'obiettivo è quello di comprendere il posizionamento di un termine all'interno della classificazione di *SentiWordNet* appena esposta.

Prendiamo in esame il termine *anxiety*; all'interno di *SentiWordNet* compare con la seguente dicitura:

PosScore	NegScore	Termine	Definizione
0.125	0.75	anxiety#2	a vague unpleasant emotion that is experienced in anticipation of some (usually ill-defined) misfortune

Il dato PosScore si riferisce al punteggio assegnato al valore *positività* e il dato NegScore, di conseguenza, è il punteggio assegnato al valore *negatività*. Come si può facilmente notare, non è presente alcun dato riguardo il grado di oggettività – soggettività della parola; questo valore viene individuato attraverso la seguente relazione:

$$ObjScore = 1 - (PosScore + NegScore)$$

Tale relazione viene fornita direttamente dal dizionario, fra i dati preliminari presentati prima di elencare i vari termini. ObjScore è coincidente con il massimo livello di *oggettività* in caso questo sia uguale a 0, mentre, viceversa, coincide con il massimo livello di *soggettività* se questo è uguale a 1.

È possibile a questo punto calcolare il valore ObjScore per il termine *anxiety*.

$$ObjScore = 1 – (0.125 + 0.75) = 0.125$$

A questo punto si dispone di tutti i dati per poter individuare la polarità del termine per entrambi gli assi, sia SO che PN; utilizzando i riferimenti dati dalla *Figura 4*, possiamo identificare il termine *anxiety* in un punto che si trovi a sud-est nella griglia triangolare.

Supponendo che *anxiety* sia solo un termine all'interno di un testo più ampio e complesso, è possibile individuare la polarità complessiva del testo procedendo con l'analisi parola per parola e poi unendo i risultati ottenuti in un valore aggregato. Il procedimento che associa ad ogni termine un valore semantico è definito *tagging* e può essere compiuto manualmente attraverso codificatori umani oppure è possibile eseguirlo attraverso l'utilizzo di *dizionari ontologici* (Ceron, Curini e Iacus, 2014), come visto in questo paragrafo.

Come ogni tecnica di analisi del *sentiment* analizzata finora, i metodi *lexicon based* non sono esenti da aspetti che ne limitano l'utilizzo e la fruibilità, come evidenziato da Ceron, Curini e Iacus (2014). Basti pensare ad esempio alla frase "che bella fregatura!"; all'interno della frase coesistono il termine "bella" e "fregatura", i quali assumono due valenze opposte in termini di

positività e negatività, facendo così risultare il valore finale della frase come neutrale, sebbene sia chiaro che la frase esprime un giudizio negativo. Il problema sussiste nella considerazione dei termini in quanto *bag of words*, come sono stati definiti in questo Capitolo, e non nel significato che assumono in base alla relazione che esiste fra di essi; se da un lato si semplifica il procedimento in fase di analisi dal punto di vista dei software e si rende loro più comprensibile la "lettura" di un testo, dall'altro si perde una parte del significato che le frasi racchiudono e il risultato può distanziarsi anche di molto dalla classificazione corretta. In casi come questi, il *tagging* manuale ridurrebbe significativamente il termine di errore della classificazione, di contro però necessiterebbe di un impiego di tempo e risorse notevolmente maggiore.

Un altro aspetto critico posto in risalto dagli autori corrisponde con la difficoltà di creare un *dizionario ontologico* aggiornato e corretto per ogni lingua; nel caso preso in analisi si è facilmente potuto analizzare la polarità di un termine in quanto la lingua inglese dispone di dizionari completi e facilmente costruibili (a causa della struttura stessa della lingua). Basti pensare, viceversa, a lingue come quelle orientali o mediorentali per rendersi conto che non in tutti i casi è facilmente costruibile un *dizionario ontologico* adeguato alle necessità richieste dalla *sentiment analysis*.

In AlOwisheq, AlHumoud, AlTwairesh e AlBuhairi (2016) viene analizzato ampiamente il problema della creazione di un *dizionario ontologico* utile ai fini della *sentiment analysis* per la lingua araba.

Il sito *SentiWordNet* e il lavoro di Agarwal e Mittal (2016) sono riferimenti aggiornati per dettagli sulla creazione di un *dizionario ontologico* e sul suo funzionamento.

A conclusione dell'analisi svolta riguardo queste due famiglie di analisi del *sentiment*, quella *machine learning* e quella *lexicon based*, si riporta, in maniera concisa, un confronto posto in essere da D'Andrea, Ferri, Grifoni e Guzzo (2015) fra le due macroaree in termini di vantaggi e limitazioni (*Tabella 4*):

Tabella 4: Confronto fra famiglie di analisi

APPROCCIO	VANTAGGI	LIMITAZIONI
Machine learning	Abilità di adattamento e di creazione di modelli *trained* per finalità specifiche e contesti particolari.	Una limitata applicabilità di nuovi dati a causa della necessità della creazione di dati classificati, operazione che potrebbero rivelarsi molto costosa, in certi casi proibitiva.

Lexicon based	Copertura più ampia dei termini.	Numero finito di parole all'interno del *lexicon* e difficoltà nell'assegnazione di un determinato orientamento semantico del sentiment di una parola, assieme al relativo punteggio.

Fonte: D'Andrea, Ferri, Grifoni e Guzzo (2015), p.29

Come si vede, la tabella evidenzia le dicotomie finora emerse: adattamento e applicabilità, copertura e costi.

2.4. Altre metodologie

Le due famiglie di metodi di analisi, di cui si è brevemente cercato di fornire una descrizione nel presente Capitolo, ovvero quella definita come *machine learning* e quella invece delle tecniche *knowledge-based*, comprendono al loro interno la maggior parte delle tecniche presenti in letteratura; tali macrogruppi non sono però riconosciuti univocamente da tutti gli autori, né tantomeno possono avere la pretesa di comprendere totalmente le tecniche disponibili al momento.

Questo a causa della natura stessa dell'ambito di studi a cui si fa riferimento, ovvero l'ambito dell'*Intelligenza Artificiale (IA)* e di come questa sia in grado o meno di comprendere il linguaggio umano e di destrutturarlo per estrapolarne il significato; essendo questo campo di studi incredibilmente fervido e dinamico, la letteratura conseguentemente segue un corso di rinnovamento costante, a volte contraddicendosi o presentando discrepanze da autore a autore.

Basti pensare a quanto è successo recentemente in casa Facebook dove alcuni sviluppatori hanno intrapreso un progetto di comunicazione fra due computer (*bot*) attraverso il linguaggio umano. Dopo aver cominciato a comunicare attraverso frasi di senso compiuto in lingua inglese, i due *bot* hanno iniziato una conversazione in una lingua completamente nuova, pur utilizzando i termini della lingua inglese. Essi, combinati in una maniera del tutto diversa da quella canonicamente prevista dalla sintassi anglosassone, producevano frasi apparentemente incomprensibili. Tale linguaggio per le due macchine costituiva invece un nuovo metodo di comunicazione totalmente coerente (Facebook ha fatto parlare tra loro due bot, e questi hanno parlato una nuova lingua, 2017). Una comunicazione fra i due *bot* è arrivata a assumere una forma del seguente tipo:

Bob: «I can can I I everything else»

Alice: «Balls have zero to me to me to me to me to me to me to me to»

Se si contestualizza l'intera analisi condotta finora in un ambito di studi come questo, dove spesso ci si trova a dover trattare tematiche al limite fra il reale e il fantascientifico, è facile comprendere come mai non ci sia un'univocità nella rappresentazione del suo funzionamento e dei suoi metodi.

Una metodologia che è necessario richiamare a questo punto dell'esposizione è quella definita come *Natural Language Processing (NLP)*. Questo approccio, differente dai metodi visti finora, si basa su tecniche di psicologia cognitiva e analisi linguistica che attraverso l'utilizzo di algoritmi permette di decodificare un testo (Ceron, Curini e Iacus, 2014). L'idea alla base di questa metodologia è quella di modellare attraverso algoritmi il modo in cui si forma il linguaggio umano; ha il pregio di essere supportata da un modello cognitivo - linguistico, ma ha il difetto di conseguenza di essere troppo legata alle assunzioni di base. Viceversa, le tecniche analizzate fino a questo punto dell'esposizione non modellano esplicitamente il modo in cui il linguaggio si forma ma cercano di trovarne una regolarità, come se si trattasse di una "scatola nera". Cercano di semplificare quanto più possibile la sovrastruttura linguistica lasciando che l'algoritmo possa apprendere sulla base del numero minore possibile di assunzioni.

Gli autori concludono, a riguardo, che questo tipo di tecniche, più che essere finalizzate all'estrapolazione vera e propria di opinioni puntano a fini esplorativi della struttura dei testi e della relazione tra parole e contenuti.

Essi delineano poi alcune altre metodologie presenti al momento nello scenario della *sentiment analysis*. Fra queste vanno annoverate le seguenti:

- *Information retrieval (IR)*; si tratta di una tecnica basata sulla ricerca di risposte a particolari domande all'interno dei documenti basandosi sull'utilizzo di alcune *keyword*.
- Si parla invece di *Information extraction (IE)* quando si cerca di estrarre una specifica informazione da un documento. Lo scopo non è quello di estrarre un'opinione quanto piuttosto la classificazione dei testi in determinate categorie.

- Si definisce *topic detection* l'insieme di tecniche atte all'identificazione o al monitoraggio dell'utilizzo di *keyword* in un *corpus* di testi che si evolve nel tempo come ad esempio siti di informazione.
- Infine vengono definite con il nome di *text summarization* quelle tecniche che cercano di sintetizzare l'informazione contenuta in un testo riconducendo quest'ultimo ad un riassunto molto contenuto che poi viene analizzato soffermandosi sul numero di volte in cui le frasi si ripetono all'interno del testo stesso e dell'insieme dei testi da analizzare ponendole in relazione con un database preesistente (Ceron, Curini e Iacus, 2014).

CAPITOLO 3. ALCUNI CASI PRATICI

3.1. Sentiment analysis e aziende

Nei precedenti capitoli è stato visto quali sono le metodologie di approccio alla *sentiment analysis* secondo le principali scuole di pensiero; si è osservato anche quali siano gli ambiti in cui questa attività è particolarmente attiva e come l'economia sia uno dei settori più rilevanti in questo senso (*Tabella 1*). In questo Capitolo ci soffermeremo in maniera più concreta su alcuni casi di applicazione pratica della *sentiment analysis*, in relazione all'attività d'azienda in primis, ma anche in relazione a campagne politiche, passando attraverso l'analisi di quelle che sono le piattaforme informatiche più utilizzate fra quelle disponibili al momento nel mercato.

3.1.1. I software di brand monitoring e la figura del social media manager

La nascita del *Web 2.0* ha portato negli ultimi anni a una serie di conseguenze (e opportunità) le quali hanno rivoluzionato in maniera sostanziale il concetto di marketing per le aziende e soprattutto le modalità di comunicazione fra aziende e consumatori. Tale rivoluzione tecnologica e questa nuova necessità da parte delle aziende di rendere più agevole, concreto e tempestivo il rapporto con i clienti hanno trovato espressione, nel mercato, in piattaforme con finalità di *brand monitoring*. Per *brand monitoring* (o *social media monitoring*, più in generale) si intende l'insieme di strumenti che consentono di analizzare la reputazione aziendale online attraverso l'ascolto delle conversazioni – *Facebook post*, *tweet*, *blog post*, *news*, ecc. – generate dagli utenti sul brand e l'analisi delle attività messe in campo (Zaccone, 2015). Tale tipo di analisi deve, secondo l'autrice, tener conto nella maniera più ampia possibile di tutte le risorse informative online, e non solo di quelle direttamente gestite dalle aziende come ad esempio pagine Facebook create ad hoc o il sito web istituzionale. I consumatori discutono fra loro scambiando opinioni riguardo l'azienda, anche (se non soprattutto) in blog e siti in cui l'azienda non ha un controllo diretto: ignorare queste fonti informative corrisponde ad accettare l'ottenimento di un feedback limitato e parziale della reputazione aziendale.

Alla luce di quanto detto finora può però sorgere la seguente domanda: quali sono le figure all'interno dell'organigramma aziendale che hanno come ruolo quello di mantenere monitorata la reputazione aziendale sul web?

Lo sviluppo dei *social* e il loro potenziale utilizzo da parte delle imprese ha dato il via alla nascita di una figura aziendale adatta a prendere le redini del *brand monitoring* e a fungere da tramite fra le esigenze dei consumatori e gli organi decisionali aziendali: il *social media manager*.

Il *social media manager* è quella figura professionale, che può essere compresa nell'organigramma dell'azienda come anche essere un consulente esterno, la quale si rivolge alle aziende, alle organizzazioni o alle istituzioni, ma anche a figure pubbliche e VIP che vogliono curare la propria immagine sui *social network*. Si occupa principalmente di realizzare una strategia di comunicazione da mettere in atto sui principali *social* come *Facebook*, *YouTube*, *Instagram*, *Twitter*, ecc. Gli obiettivi di un *social media manager* possono essere diversi, ad esempio migliorare la *brand awareness*, l'immagine di una azienda, aumentare le vendite di un prodotto e altro (Wikipedia, 2017).

Sono queste figure professionali, in primis, ad approcciarsi, nel mondo dell'azienda, agli strumenti di *sentiment analysis* esposti in questo elaborato. Sono molte le piattaforme che rendono più agevole il lavoro dei *social media manager*; in questo paragrafo verranno citati alcuni esempi fra i più illustri, senza alcuna pretesa di fornire una panoramica completa del mercato degli strumenti di *brand monitoring* ma piuttosto avendo come finalità quella di dare un'idea generale di come la *sentiment analysis* trovi applicazione nel mondo aziendale.

Di seguito, i principali mezzi informatici per il monitoraggio della reputazione aziendale.

- *Brandwatch*: è una fra le piattaforme di *social monitoring* più utilizzate al mondo. Accedendo al sito internet del software, è possibile verificarne le aziende partner; fra queste compaiono nomi illustri quali *Fox*, *Sky*, *The Economist*, *ESPN*, *Kellog's* e *Walmart*. Di quest'ultima azienda è possibile analizzare il caso di studio, che delinea gli obiettivi strategici dell'impresa nei confronti dell'ambito *social* e di come questi obiettivi siano stati raggiunti. Attraverso un'intervista a Chandler Wilson, Director of Analytics and Insights di Walmart, si può scoprire molto sull'organizzazione aziendale dell'impresa leader della distribuzione al dettaglio e di come questa si sia modellata secondo le esigenze dei clienti e secondo l'evoluzione della tecnologia. Wilson spiega come Walmart abbia rivoluzionato il proprio approccio riguardo l'acquisizione dei dati online: *"Stiamo creando un network strutturato per key people e key themes, e attraverso il confronto con dati economici e statistici cerchiamo di costruire una strategia ad alta risoluzione"*. Walmart, con il suo peso in termini sia

economici che di impatto sociale, ha utilizzato i *social* anche per far conoscere alla popolazione americana il proprio disaccordo nei confronti delle politiche salariali federali dello Stato dell'Arkansas fissando un salario minimo aziendale maggiore del salario minimo federale. Si è in seguito potuto verificare come questa manovra abbia causato dei sobbalzi nell'andamento del mercato delle *commodities* e come abbia influenzato allo stesso modo anche l'andamento dei tassi d'interesse.

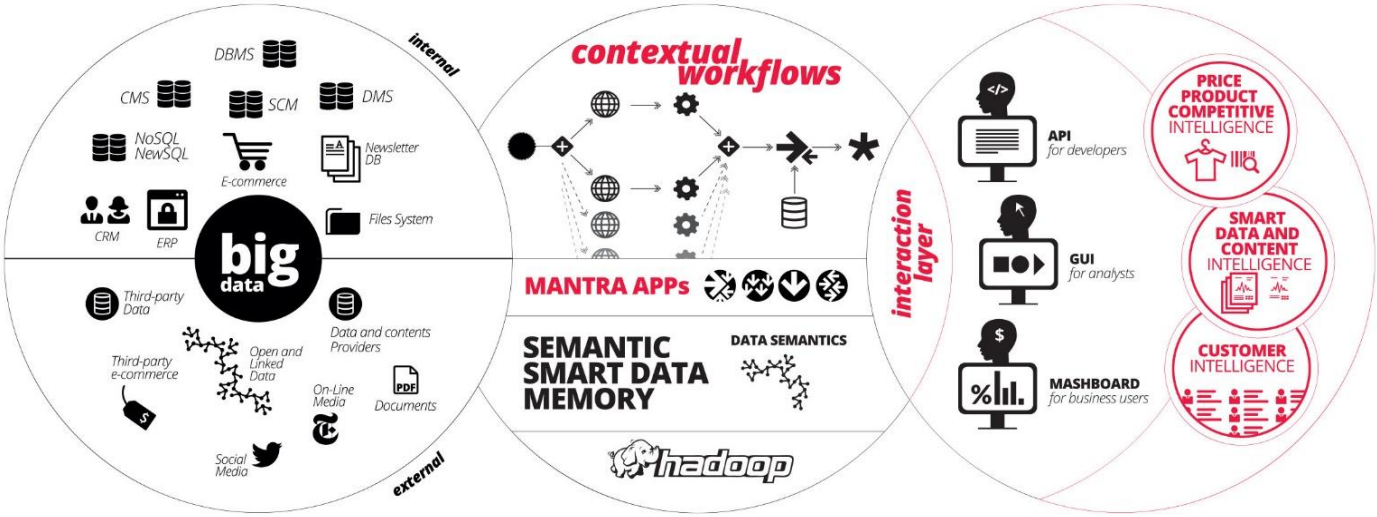
- *Lithium*: anch'essa è una piattaforma che si pone l'obiettivo di controllare a 360 gradi il gradimento e la *brand reputation* delle imprese online. Fra i *customer* più illustri si possono annoverare *HP*, *Sony*, *Best Buy*, *Virgin*, *Symantec*, *British Gas* e *Deutsche Telekom.* Viene presa in analisi la testimonianza di Kriti Kapoor, Global Director della divisione Social Customer Care di HP: *"Il nostro successo è dovuto a una strategia composta da tre fasi. Primo, noi ascoltiamo. Prestiamo attenzione alle conversazioni e ai feedback nei canali social per scovare i diversi customer needs. Secondo, noi coinvolgiamo. Disponiamo di un supporto peer-to-peer e di figure dedicate che assicurano che alle domande poste dai clienti vengano date risposte immediate e vengano indirizzate nelle mani dei professionisti più esperti. Terzo, usiamo Lithium Social Intelligence (LSI) e dei report esecutivi per imparare e incanalare quanto appreso nella progettazione dei prodotti e nella nostra cutomer care strategy."*

- *Mantra:* questo software creato da Altilia è il fiore all'occhiello delle *Smart Data Platforms* made in Italy. Il suo obiettivo è quello di convertire *Big Data* in *Smart Data,* ovvero di ottenere e riorganizzare i dati in modo tale che la loro lettura sia conveniente con le finalità preposte dall'utilizzatore. Il raggiungimento di questo obiettivo passa attraverso algoritmi basati su tecniche che abbiamo visto nel Capitolo 2, come ad esempio sistemi *machine learning* e metodi proposti dalla *natural processing language*. Nella *Figura 5* viene data una rappresentazione schematica di come funziona il processo di raccolta di informazioni e di come queste vengano presentate al cliente; attraverso il confronto fra dati ottenuti internamente e dati esterni, Mantra elabora e processa l'intero bagaglio informativo, per riorganizzarlo infine in *insights* per gli sviluppatori, i manager e gli analisti dell'impresa.

- A conclusione di questa breve rassegna non si possono non citare le soluzioni proposte dalle stesse piattaforme social in cui solitamente si opera per sviluppare un'analisi del *sentiment*. Per fare qualche esempio, basti aprire la homepage del

motore di ricerca Google: si può notare che, oltre alla canonica barra di ricerca, compare la scelta *Soluzioni aziendali*; aprendo questa sezione, le aziende possono accedere a una serie di strumenti che sfruttano il sistema di ricerca per *keywords* di Google offrendo all'utilizzatore la possibilità di proporsi in maniera più efficace ai consumatori oltre che a tenere sotto stretto controllo il modo e la portata con cui l'azienda stessa viene conosciuta attraverso il web. Anche Facebook e Twitter propongono simili soluzioni, alcune di esse mirate ad una maggior visibilità ed altre invece finalizzate alla pubblicazione programmata relativamente di *post* e di *tweet*, seguiti da un controllo dell'ampiezza dei contatti raggiunti e dalla possibilità di rispondere in tempo reale ai commenti dei consumatori.

Figura 5: Fasi del lavoro in Mantra



Fonte: http://www.altiliagroup.com/platform/mantra-platform

3.1.2. Monitoring di una campagna di marketing attraverso Twitter e il ruolo degli influencer

Dopo aver visto quali sono le principali piattaforme offerte al pubblico, e di come queste vengano utilizzate dalle aziende, si procede con l'esposizione di un esempio dell'applicazione delle tecniche di *sentiment analysis* e di *opinion mining* fin qui analizzate.

Benedetto e Tedeschi (2016) presentano un'analisi di alcune campagne di *brand monitoring* riferite ai *tweet* raccolti relativi a un noto brand operante nel settore tecnologico (il quale rimarrà anonimo) e più in particolare riguardo l'evento del lancio di una nuova serie di *smartphone* sul mercato. L'analisi condotta dagli autori verrà in seguito riportata attraverso i suoi punti salienti,

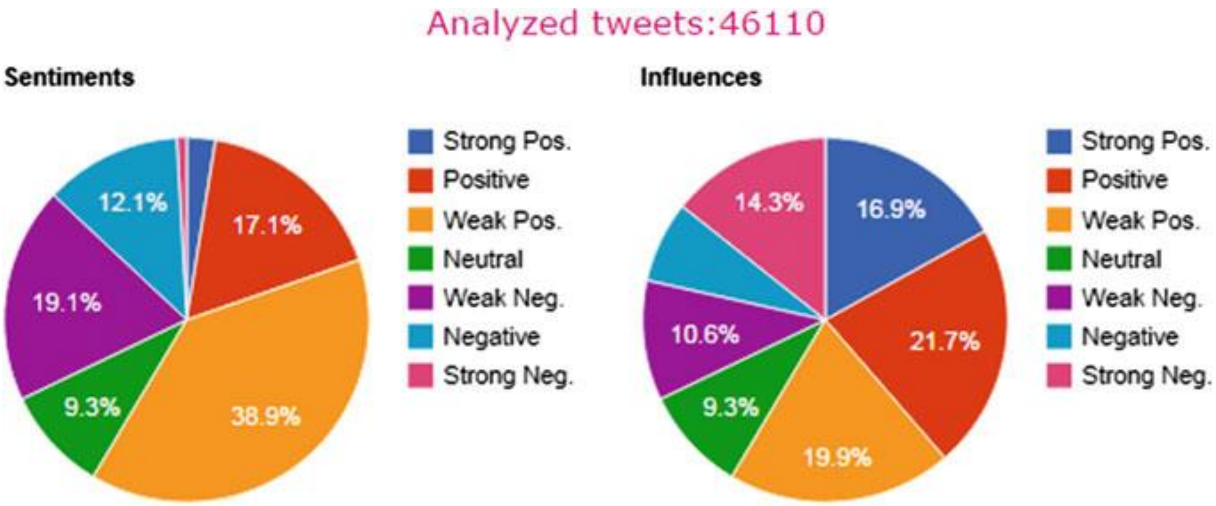
i dettagli e una più ampia spiegazione essendo disponibile nel rapporto completo redatto dagli autori.

L'analisi muove i suoi assunti sulla base di dati raccolti su Twitter (46110 *tweet* analizzati). Attraverso il confronto e la fusione dei dati ottenuti in vari *data set* gli autori individuano la polarità dei *tweet* analizzati e raffigurano i risultati in alcuni grafici in grado di riassumere in maniera più chiara il loro significato.

La *Figura 6* presenta due grafici a torta: sulla sinistra troviamo la classificazione dei *tweet* in base al loro *sentiment* in valore assoluto (definendo sette gradi di polarità, ordinati da *strong negative* a *strong positive*) esprimendo i dati in percentuale sul totale.

Come evidenziato nel grafico, il 38,9% dei *tweet* può essere raggruppato sotto la polarità *weak positive* mentre il 19,1% dei *tweet* rientra nella categoria *weak negative*; si definisce così per la maggior parte dei *tweet* un sentiment a favore dell'azienda; a destra invece viene rappresentata una classificazione non solo in base alla polarità dei *tweet* ma anche in base alla *popolarità* dell'autore del *tweet*; in questo caso si può notare come le categorie *strong positive* e *strong negative* raggiungano dei valori molto più sostanziosi. Le figure degli *influencer* giocano dunque un ruolo molto importante nel determinare le preferenze degli utenti sul web e possono in alcuni casi modificare notevolmente i dati recepiti dalle aziende.

Figura 6: Sentiment (sinistra) e Influenza (destra) suddivisi in sette gradi di polarità.



Fonte: Benedetto e Tedeschi (2016), p.366

La popolarità di chi scrive e pubblica un *tweet* è dunque un elemento chiave nell'analisi del problema, aumentando ulteriormente il coefficiente di difficoltà per gli analisti, in quanto viene aggiunta al sistema una variabile rilevante. Non a caso le imprese, per aumentare la propria

notorietà e accrescere la propria reputazione, decidono sempre più spesso di affidarsi agli *influencer* del mondo del web; il proliferare di collaborazioni fra aziende e personaggi di rilievo in piattaforme come *YouTube*, con collaborazioni fra *influencer* e case videoludiche, o *Instagram*, in cui si vedono sempre più spesso personaggi famosi affiancati da prodotti (o più semplicemente al nome) dei brand più disparati ne costituiscono degli esempi lampanti. Basti osservare la campagna *social* di Carrera, celebre brand produttore di occhiali, il quale dispone di una partnership con il noto attore e musicista Jared Leto; non sempre è però necessario affiancare il nome del brand a personaggi di fama internazionale in quanto a seconda del mercato di riferimento è fondamentale individuare quelle figure il cui parere è in grado, appunto, di influenzare l'opinione pubblica.

3.2. Sentiment analysis e politica

Abbiamo visto come lo studio del *sentiment* sia applicato in larga misura nel contesto aziendale, attraverso l'utilizzo di piattaforme di *brand monitoring* in grado di fornire informazioni utili agli analisti e al management dell'azienda.

Come si è evidenziato nel Capitolo 1 il mondo aziendale è solo uno fra i contesti in cui la *sentiment analysis* si trova ad operare: la politica, ambiente ideologico in cui la democrazia e la libera opinione trovano la loro massima espressione, è un altro ambito di ricerca in cui gli studiosi del *sentiment* si trovano ad operare in maniera concreta, al fine di prevedere, analizzare e individuare le preferenze nei confronti di partiti politici, referendum, proposte di legge.

Nel contesto politico i sondaggi sono il mezzo principe attraverso il quale i media, sia *offline* che *online*, giudicano il livello di gradimento; per questo motivo la raccolta dei dati ai fini della creazione dei sondaggi è spesso e volentieri l'unica attività, anche per gli stessi partiti politici, in grado di fornire un riscontro effettivo delle scelte compiute in sede di campagna elettorale. A questi dati negli ultimi tempi si è affiancata la *sentiment analysis* come strumento integrativo ai mezzi più canonici di ricerca di informazioni. Questo andamento è giustificato dallo sviluppo e dalla conseguente rilevanza crescente nel contesto informativo ma non solo. Anche il costo relativamente contenuto della *sentiment analysis* rispetto a quello associato alla costruzione di un sondaggio nella maniera classica, nonché dalla maggior reattività in tempo reale a seguito degli avvenimenti che si susseguono durante la campagna, sono fattori promettenti: in un sondaggio, dalla raccolta dei dati alla loro rappresentazione possono passare dei giorni, mentre con strumenti di monitoraggio online si può analizzare ora dopo ora la variazione in termini di gradimento senza dover incorrere in *delay* temporali eccessivi. Inoltre, l'utilizzo di tecniche di *sentiment analysis* in ambito elettorale permette (ed è uno dei pochi casi in cui questo è

possibile) di parlare di vero e proprio *forecasting*: se nell'ambito aziendale le performances possono essere soggette ad una distorsione dovuta alla soggettività di alcuni dei dati che confluiscono nei bilanci, nel mondo politico i risultati vengono inconfutabilmente resi pubblici attraverso i voti dei cittadini ed è quindi possibile confrontare i dati previsionali raccolti durante la campagna elettorale con i dati effettivi al termine del conteggio dei voti raccolti alle urne (Ceron, Curini e Iacus, 2014).

Una fra le piattaforme online più utilizzate per lo studio del gradimento politico, come anche del livello di felicità (*iHappiness*), attraverso l'utilizzo delle tecniche di *sentiment analysis*, è *Voices from the Blogs (VfB)*: si tratta di un a piattaforma che analizza il *sentiment* espresso sui social media, blog e web utilizzando avanzate metodologie statistiche proprietarie disegnate per la *sentiment analysis*. L'approccio di VfB permette di interpretare e sintetizzare con precisione statistica i *Big Data* velocemente e indipendentemente dalla lingua di origine dei testi. (Voices from the Blogs S.r.l. è uno Spin-off Università degli Studi di Milano. Società iscritta al Registro delle Startup Innovative, fondata il 12 dicembre 2012).

Questa tematica viene ampiamente approfondita da Ceron, Curini e Iacus (2014) (fondatori di Voices from the Blogs), attraverso lo studio dei risultati ottenuti tramite l'ascolto della voce degli elettori sul web e il raffronto fra i dati ottenuti e i risultati pubblicati dalle testate giornalistiche, analizzando i punti di forza e le criticità dell'utilizzo delle tecniche di *sentiment analysis* (in particolare *iSA*). Viene di seguito riportato un caso pratico pubblicato dagli autori, sintetizzato e trattato attraverso i passaggi più rilevanti ai fini dell'esposizione.

3.2.1. Le primarie del centrosinistra, 2012

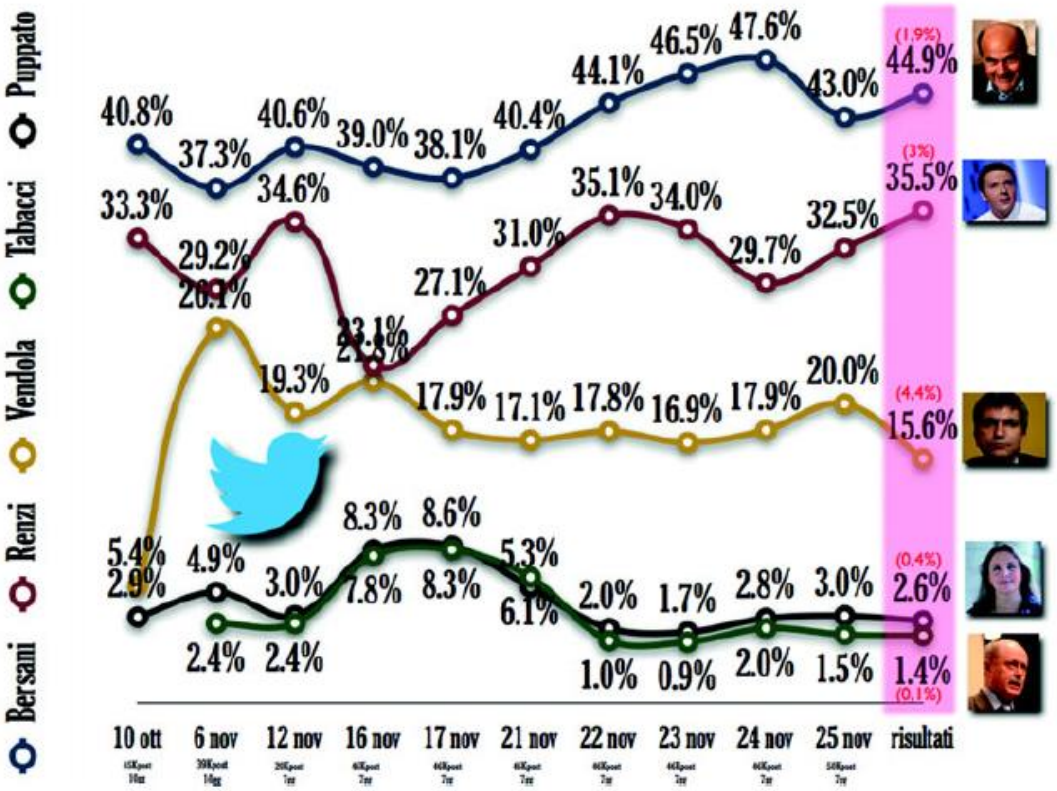
Nel 2012 gli elettori italiani sono stati chiamati a eleggere il loro rappresentante nella coalizione politica di centrosinistra *Italia. Bene Comune*, in vista delle elezioni politiche del 2013. Il sistema prevedeva un doppio turno; al termine del secondo turno si sarebbe deciso l'esponente politico rappresentante della coalizione. A presentarsi come candidati a questo ruolo sono stati cinque personaggi della scena politica del centrosinistra: Pierluigi Bersani (PD), Matteo Renzi (PD), Nichi Vendola (SEL), Laura Puppato (PD) e Bruno Tabacci (ApI). Gli autori dello studio hanno raccolto più di 600.000 *tweet* in un periodo compreso fra il 6 ottobre e il 2 dicembre 2012, analizzando giorno per giorno l'andamento delle preferenze online.

Prima di procedere con l'esposizione dei dati è utile fare una doverosa precisazione. Essa risiede nella differenza sostanziale fra i candidati in termini di popolarità online: se da una parte un candidato come Bersani godeva del supporto del partito in maniera più ampia, e presentava

già dalle prime battute un margine di preferenza maggiore rispetto a quello degli altri esponenti politici, Renzi al contrario era (ed è tuttora) una figura di maggior rilievo online, nonostante i primi sondaggi lo vedessero in svantaggio rispetto al collega del PD. Se ci si fosse basati meramente sul conteggio delle menzioni di Bersani e Renzi, si sarebbe potuto notare come i due candidati godessero della stessa percentuale di citazioni positive, ma Renzi risultava più "popolare" in quanto maggiormente menzionato fra le conversazioni inerenti sia il primo che il secondo turno.

Gli autori di *Voices from the Blogs* sintetizzano l'andamento e le fluttuazioni giorno per giorno delle preferenze nella *Figura 7* raccogliendo per ognuno dei valori rappresentato sulle ascisse dai 40.000 ai 50.000 tweet. L'ultima colonna rappresenta i dati reali ottenuti alle urne, evidenziando lo scarto dai dati previsionali.

Figura 7: Fluttuazione delle preferenze di voto dei candidati



Fonte: Ceron, Curini e Iacus (2014), p.96

Come si può facilmente osservare, i dati assestavano in vetta alla classifica Bersani con un punteggio del 43%, seguito da Renzi (32,5%); gli scarti fra i dati effettivi e quelli previsionali sono ridotti, raggiungendo il loro picco massimo nel caso di Vendola (4,4% di scarto), avvalorando così le doti di *forecasting* degli strumenti utilizzati.

Nella fase del secondo turno vengono analizzati quasi 25.000 tweet pubblicati fra il 29 novembre e il 1 dicembre, il giorno precedente alle elezioni. Durante la seconda tornata alle urne, la scelta degli elettori poteva ricadere fra i due candidati che avevano ottenuto più voti, ovvero i due esponenti del PD Bersani e Renzi. Nella *Tabella 5* viene esposta la situazione al termine del secondo turno in raffronto alle stime previsionali eseguite da *Voices from the Blogs* (sotto la voce *iSA*) e dai principali sondaggi. I risultati decretano Bersani vincente con più del 60% delle preferenze; a seguire Renzi con un punteggio di 39,1%. Come si può evincere osservando i dati riportati nella tabella, la tecnica di *sentiment analysis* messa in atto dagli autori è stata, fra i metodi previsionali, la più accurata: con i dati previsti di 58,4% per Bersani (scarto di 2,5 punti in difetto rispetto ai dati effettivi) e di 41,6% per Renzi (scarto di 2,5 punti in eccesso rispetto ai dati effettivi) è la stima che meglio rappresenta le preferenze degli elettori in termini di gradimento.

Tabella 5: Confronto dei dati nel secondo turno

	Giorno	*Bersani (%)*	*Renzi (%)*	*Gap Bersani-Renzi*
Voti reali	-	60,9	39,1	+21,8
iSA	01/12/2012	58,4	41,6	+16,8
Ipsos	29/11/2012	57,5	42,5	+15
Quorum	28/11/2012	56,4	43,6	+12,8
SWG	28/11/2012	55	45	+10
COESIS	28/11/2012	54	46	+8
ISPO	27/11/2012	56,5	43,5	+13
IPR	26/11/2012	56	44	+12
PIEPOLI	25/11/2012	59	41	+18

Fonte: Ceron, Curini e Iacus (2014), p.98

Concludendo, gli autori affermano che nonostante i limiti e le problematiche legati all'analisi del *sentiment*, nel campo delle previsioni elettorali, i risultati forniscono delle ragioni per essere ottimisti riguardo le potenzialità e le possibilità che questo campo di studi offre fornendo un prezioso strumento di supporto alle tecniche tradizionali di sondaggio demoscopico *offline* (Ceron, Curini e Iacus, 2014).

CONCLUSIONI

Alla luce di quanto messo in evidenza attraverso i Capitoli dell'elaborato, si può apprezzare come lo sviluppo della competitività *online*, scoppiata a causa di una maggior consapevolezza da parte delle imprese dell'importanza del ruolo dei *social* nelle vite dei consumatori, abbia dato vita alla crescita sostenuta di un gran numero di tecniche di *sentiment analysis*. Abbiamo visto come queste tecniche, se applicate tenendo in considerazione alcuni fattori, possano fornire risultati utili, come nei casi visti: nel campo aziendale attraverso la raccolta di informazioni utilizzabili per migliorare i prodotti e i servizi offerti ai clienti, nell'ambito politico attraverso il costante monitoraggio delle preferenze in una campagna elettorale. Si è inoltre sottolineato che, per ottenere dei buoni risultati, è necessaria una stretta collaborazione fra la potenza di calcolo dei computer e le capacità interpretative dell'uomo; come affermato da Nate Silver, famoso statistico americano, *"the key to making a good forecast is not in limiting yourself to quantitative information"*.

Portando a termine il ragionamento svolto finora è interessante notare come non sia possibile mettere un punto fermo a questo scenario. Si possono certamente individuare alcuni ambiti, in parte esplorati, che lasciano spazio alla crescita di applicazioni pratiche delle metodologie di analisi del *sentiment* congiuntamente all'apertura di nuovi orizzonti per il loro utilizzo.

Sintetizzando, se ne propongono almeno tre.

Dal punto di vista della ricerca, in primo luogo, c'è un ampio margine di miglioramento riguardo i limiti precedentemente analizzati; ci si riferisce, soprattutto, a quegli ostacoli linguistici che si frappongono fra i testi e le capacità interpretative dei computer, sebbene l'avanzamento tecnologico stia lavorando per appianare tali ostacoli e rendere più vicini i mondi della semantica e dell'informatica.

Secondariamente, sotto un'ottica economico-aziendale, l'applicazione delle tecniche di *sentiment analysis* risulta essere relativamente costosa e, conseguentemente, impiegata in larga scala quasi esclusivamente dalle grandi aziende le quali possono più agevolmente decidere di adibire parte del loro budget a *sistemi informativi di marketing* in grado di svolgere le operazioni inerenti l'analisi del *sentiment*. In un contesto nazionale, costituito per la stragrande

maggioranza da piccole e medie imprese, l'applicazione di tali tecniche può riscontrare alcune difficoltà e risultare carente rispetto alle potenzialità.

Infine, osservando l'argomento nella logica della definizione di un organigramma aziendale, si pongono delle prospettive di sviluppo di figure quali il *social media manager* (come visto nel Capitolo 3); questa figura professionale, unitamente allo sviluppo di canali di distribuzione *online* (*e-commerce*), sta assumendo un ruolo sempre più rilevante nel contesto aziendale e, dal punto di vista delle risorse umane, sempre più aziende stanno integrando questa figura all'interno del proprio organigramma.

BIBLIOGRAFIA

- Agarwal, B. e Mittal, N. *Prominent Feature Extraction for Sentiment Analysis.* Springer, 2016.
- AlOwisheq, A., AlHumoud, S., AlTwairesh, N. e AlBuhairi, T. «Arabic Sentiment Analysis Resources: A Survey.» In *Social computing and Social Media*, di Meiselwitz, G., 267-277. Springer, 2016.
- Benedetto, F. e Tedeschi, A. «Big Data Sentiment Analysis for Brand Monitoring in Social Media Streams by Cloud Computing.» In *Sentiment Analysis and Ontology Engineering*, di Shyi-Ming, C. e Witold, P. 341-377. Springer, 2016.
- Bonzanini, M. «Stemming, Lemmatisation and POS-tagging with Python and NLTK.» *Marco Bonzanini.* 26 Gennaio 2015. https://marcobonzanini.com/2015/01/26/stemming-lemmatisation-and-pos-tagging-with-python-and-nltk/.
- Brownlee, J. «Machine Learning Mastery.» *Supervised and Unsupervised Machine Learning Algorithms.* 16 Marzo 2016. http://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/.
- Ceron, A., Curini, L. e Iacus, S.M. *Social Media e Sentiment Analysis, L'evoluzione dei fenomeni sociali attraverso la rete.* Springer, 2014.
- D'Andrea, A., Ferri, F., Grifoni, P. e Guzzo, T. Approaches, Tools and Applications for Sentiment Analysis Implementation In *International Journal of Computer Applications (0975 – 8887) Volume 125 – No.3, September 2015,* 26 -33. 2015
- Esuli, A. e Sebastiani, F. SentiWordNet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC 2006,* 417-422. 2006
- «Facebook ha fatto parlare tra loro due bot, e questi hanno parlato una nuova lingua.» *www.ilpost.it.* 1 Agosto 2017. http://www.ilpost.it/2017/08/01/intelligenza-artificiale-inventare-nuovi-linguaggi/.
- Farhadloo, M. e Rolland, E. «Fundamentals of Sentiment Analysis and Its Applications.» In *Sentiment Analysis and Ontology Engineering*, di Shyi-Ming, C. e Witold, P. 1-24. Springer, 2016.
- Haddi, E., Liu, X. e Shi, Y. «The Role of Text Pre-processing in Sentiment Analysis.» *Sience Direct.* 2013. http://www.sciencedirect.com/science/article/pii/S1877050913001385.
- *Internet World Stats.* 13 Luglio 2017. http://www.internetworldstats.com/stats.htm.
- Kotler, P., Armstrong, G., Ancarani, F. e Costabile, M. *Principi di marketing.* Pearson, 2015.
- Liu, B. *Sentiment Analysis and Opinion Mining.* Morgan & Claypool Publishers, 2012.

- Medhat, W., Hassan, A. e Korashy, H. «Sentiment analysis algorithms and applications: A survey.» *Sience Direct.* Dicembre 2014. http://www.sciencedirect.com/science/article/pii/S2090447914000550.
- Scott, J. *Social Network Analysis, A Handbook.* Londra: SAGE Publications, 2000.
- *Wikipedia.* 12 Luglio 2017. https://it.wikipedia.org/wiki/Social_media.
- *Wikipedia.* 30 Aprile 2017. https://it.wikipedia.org/wiki/Social_media_marketing.
- Zaccone, E. «Web in Testa.» *Cosa è il monitoraggio dei Social Media e come organizzarlo.* 4 Giugno 2015. https://www.webintesta.it/monitoraggio-dei-social-media-a-cosa-serve-come-organizzarlo/.
- Zarella, D. *The Social Media Marketing Book.* Sebastopoli: O'Reilly Media, 2009.